Алгоритм семантического поиска в больших текстовых коллекциях

Виталий Савченко

АлтГТУ им. И. И. Ползунова, Барнаул, Россия
64svv@rambler.ru

Аннотация В статье рассматривается метод семантического поиска для многопоточной обработки текстов большого объема. Поисковый запрос и обрабатываемый текст преобразуются в графы семантических связей. Предлагается алгоритм вычисления коэффициента соответствия семантических графов. Приводятся оценки времени обработки.

Ключевые слова: семантический анализатор, граф, справочник, вес, тип семантической связи.

1 Общие сведения

В наше время, в условиях большого и стремительно растущего объема информации, актуальна задача поиска в больших текстовых коллекциях [1]. Одним из вариантов поиска является семантический поиск, т.е. поиск с точки зрения содержащейся в тексте информации [2,3,4]. Среди наиболее популярных систем семантического поиска можно выделить Google, SearchMonkey, Freebase и AskNet. Однако они имеют определенные недостатки, такие как: применение семантики лишь для незначительного улучшения результатов поиска, ограничение на длину запроса, снижение качества поиска с увеличением поискового запроса. Кроме того большинство из таких поисковых систем работают только с английским языком.

Учитывая сложность семантического поиска необходимо применять методы, основанные на имеющихся в системе знаниях о предметной области.

2 Семантический поиск

В данной работе представлен результат разработки системы семантического поиска для больших текстовых коллекций на русском языке. Ключевой особенностью полученной системы - является снятие ограничений на величину поискового запроса и многопоточная обработка текстовой коллекции.

Исходными данными для поиска являются текстовые коллекции и запрос пользователя, который представляет собой текстовую коллекцию. Исходя из

предположения, что большая текстовая коллекция в общем случае неоднородна и с точки зрения поиска интересна ее определенная часть, то текст нужно разделить на определенные участки - страницы, абзацы или наборы из нескольких предложений. Такие фрагменты будем называть «окнами».

Для каждого окна запроса и поисковых коллекций строится граф семантических связей, назовем его «семантический граф». Семантический граф представляет собой направленный граф, вершинами которого являются слова русского языка, представленные в нормальной форме, а ребра характеризуются весом и типом семантической связи. Направление ребра зависит от типа семантической связи, например, отношение объект - действие, объект - свойство, действие - время.

Для построения семантического графа каждое предложение из окна коллекции обрабатывается семантическим анализатором. В данной работе используется семантический анализатор RML[1].

Предложения окна обрабатываются последовательно. На каждой итерации семантический граф предыдущей итерации объединяется с графом G_{new} обрабатываемого предложения. Веса ребер семантического графа G_{new} равны 1. После объединения у графа G_{i+1} все веса ребер умножаются на коэффициент затухания η.

$$G_{i+1} = (G_i + G_{new}) * \eta \tag{1}$$

После этого результирующий семантический граф используется для следующей итерации. Затем из графа удаляются ребра с весом меньше δ. Уменьшение веса ребер на заданный процент, аналогичное испарению феромона в «муравьином алгоритме» [5], сделано для ослабления воздействие предшествующих семантических зависимостей между вершинами графа.

Далее необходимо подсчитать величину коэффициента соответствия семантического графа запроса и семантического графа окна. Простой поиск наибольшего общего подграфа, даже с учетом совпадения не только вершин, но и типов ребер, не приведет к цели. Во-первых, по причине NP- полноты данной задачи. Во-вторых, один и тот же смысл содержится в текстах разного стилевого оформления, например, содержит обобщающие сведения или только частичную информацию. К хордовым, обитающим в тайге, в том числе относятся и зайцы тайги. Очевидно, улавливается связь: хордовые $\rightarrow$ зайцы. Однако между хордовыми и зайцами не должно быть полного отождествления, т. к. хордовые – это не только зайцы.

Для поиска связанных по смыслу слов был использован словарь, в котором представлен перечень слов в нормальной форме [6]. Каждому слову сопоставлен набор слов, связанных с ним ассоциативной, синонимичной и т. д. связью. Таким образом, словарь представляет собой направленный граф $G_{word} = (V_{word}, U_{word})$, где вершины V_{word} - это слова в нормальной форме, а ребра U_{word} имеют действительные весовые коэффициенты от 0 до 1. Назовем граф G_{word} графом справочника.

[1] `http://www.aot.ru`

За коэффициент связанности слов a_k и a_m - вершин семантического графа запроса $G_{request}$ и семантического графа окна коллекции G_{text} возьмем произведение весов от таких же слов до общего предка в графе справочника. При совпадении слов данный коэффициент будет равен 1, иначе будет принадлежать промежутку [0;1].

Замечание: Граф справочника является упрощенной моделью знаний о реальном мире, а общий предок слов a_k и a_m в этом графе - это некоторое обобщение соответствующих понятий. Нет смысла искать общего предка слов во всем графе справочника. Следовательно, нас интересует некое ξ окружение искомых слов. В противном случае считаем, что слова никак не связаны по смыслу. Фрагмент графа G_{word} представлен на рис. 1.

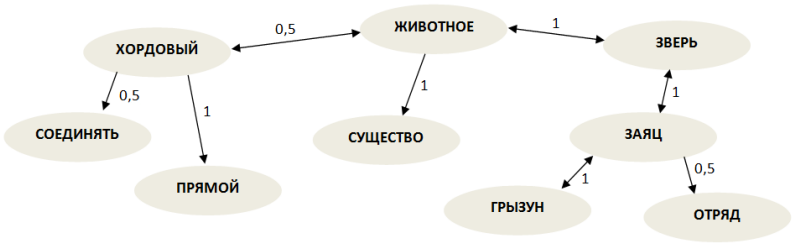

Рис. 1. Граф справочника

Далее ищем наилучшее совпадение семантического графа запроса с графом окна. Коэффициент соответствия рассчитываем по формуле 2:

$$S = \sum_{k=1}^{n} D1_k * D2_k * L1_k * L2_k, \tag{2}$$

где n – количество совпавших ребер графа запроса и окна, k – индекс соответствующего ребра, $D1_k$, $D2_k$ – произведение весов ребер в графе справочника от слова запроса и слова окна до общего предка соответственно, $L1_k$ – вес ребра k семантического графа запроса, $L2_k$ – вес ребра k семантического графа окна. Так как вариантов совпадения графов много – нас интересует максимальное значение коэффициента соответствия. Определив максимальное значение по всем окнам текстовой коллекции - получим общее значение – коэффициент соответствия запроса и текстовой коллекции.

3 Тесты и результаты

Оценка полученной системы является экспертной. Для поиска были отобраны текстовые коллекции большого объема удовлетворяющие одному и тому же запросу к поисковой системе google.ru. В зависимости от настроек

системы были получены различные значения коэффициента соответствия между поисковым запросом и коллекцией. Однако, для коллекций по содержанию которых строился запрос или коллекций аналогичного содержания значение коэффициента минимум на порядок превосходило значение коэффициента других коллекции, не похожих по содержанию.

Временные затраты на обработку коллекции в зависимости от количества предложений в запросе и окне коллекции, а так же относительные временные затраты при многопоточной обработке представлены на рис. 2.

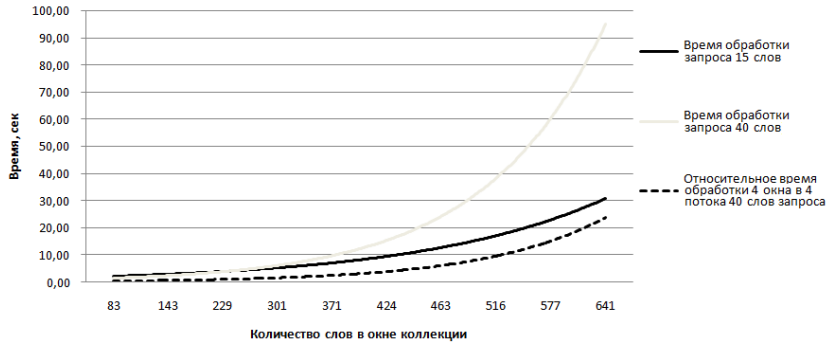

Рис. 2. Временные затраты

Основным минусом текущей реализации алгоритма является значительное увеличение времени обработки с ростом размеров окна и запроса. Плюсом является то, что текстовая коллекция и запрос могут быть разделены на окна оптимального размера с точки зрения времени обработки. Кроме того обработка окон текстовой коллекции, в данном случае, может выполняться параллельно, что значительно повышает скорость выполнения на многопоточных и многопроцессорных системах.

Список литературы

1. *Hannah Bast, Marjan Celikik* Efficient Fuzzy Search in Large Text Collections // ACM Transactions on Information Systems, 2010.
2. *Mathieu d'Aquin, Enrico Motta* Watson, more than a Semantic Web search engine // IOS Press Amsterdam, 2011.
3. *K Elbedweihy, S N Wrigley, F Ciravegna, D Reinhard, A Bernstein* Evaluating Semantic Search Systems to Identify Future Directions of Research // Second International Workshop on Evaluation of Semantic Technologies, page 25-36, 2012.
4. *G. Tsoumakas, M. Laliotis, N. Markantonatos, I. Vlahavas* Large-Scale Semantic Indexing of Biomedical Publications at BioASQ // BioASQ Workshop, 2013.
5. *Штовба С. Д.* Муравьиные алгоритмы // Exponenta Pro. Математика в приложениях, №4, с.70-75, 2003

6. *Крайванова В.А., Кротова А.О., Крючкова Е.Н.* Построение взвешенного лексикона на основе лингвистических словарей // Материалы Всероссийской конференции с международным участием ЗОНТ-2011, Т.2, Новосибирск, 2011.

Semantic Search Algorithms in Large Text Collections

Vitaliy V. Savchenko

Altai State Technical University, Barnaul, Russia
64svv@rambler.ru

Abstract. This article describes a method of semantic search based on the text processing of large volume. Search requests and processing text from analyzed collection is transformed into a graph of semantic relationships, the comparison of which allows us to define a measure of semantic similarity of compared texts. An algorithm is proposed to calculate the coefficient of semantic graphs concordance. Estimates of the processing time are also given.

Keywords: semantic analyzer, graph, directory, weight, type of semantic communication.